Perception of Pitch Tails at Potential Turn Boundaries in Swedish

Margaret Zellers

Department of Speech, Music & Hearing, Kungliga Tekniska Högskolan, Stockholm, Sweden

zellers@kth.se

Abstract

In a number of languages, intonational patterns at prosodic boundaries are considered to be relevant for turn transition or turn hold. A perception experiment tested the influence of fundamental frequency (F0) peak height and rising final contours on Swedish listeners' judgment about whether a speaker wanted to hold the turn. While F0 peak height, as has been previously shown, did influence listeners' judgments, the end height of rising pitch tails apparently did not influence listeners' judgments about whether a speaker planned to continue talking, even though they showed sensitivity to the differences in a discrimination task. The differences in responses in the tasks, as well as the difference from results found for other languages, may indicate that listeners used comparative prominence to guide their judgments, rather than intonation playing a direct role in the turn-transition system.

Index Terms: prosody, perception, turn-taking

1. Introduction

Intonational patterns at prosodic boundaries have been associated with cuing turn hold or transition. High pitch rises or low falls tend to occur at turn transitions in Tyneside English, along with slowing and vowel centralization (Local, Kelly & Wells [1]). Gravano & Hirschberg [2] report similar intonational features in a corpus of American English. Caspers [3] reports that in Dutch, level intonation and to an extent rising intonation (i.e. % and H% boundary tones, respectively) are associated with turn hold. Koiso et al. [4] report that moving pitch contours at prosodic boundaries are associated with turn hold in Japanese, and furthermore, that no intonational features at these locations are neutral to the turn-hold/turn-change contrast.

In Central Swedish, variation in phrase-final pitch may be a more complicated question, as words carry obligatory contrastive pitch accents, and the final pitch accent is often followed by an additional H- tone marking focus (cf. Bruce [5], Gårding [6]; Gussenhoven & Bruce [7]). House [8, 9] demonstrated that rising intonation, while acceptable as a means of marking questions in Swedish, occurs in only 22% of cases, with children more likely than adults to use this type of marking. This suggests that manipulation of boundary material may not be a preferred method of signaling meaning in Swedish, even if it is permissible.

Hjalmarsson [10] reports that in her Swedish data, "flat" intonation was associated with turn-holding, while "falling" intonation was associated with turn-yielding. This is consistent with findings by Edlund & Heldner [11], who also found that rising intonation patterns were not consistently associated with either turn-holding or turn-yielding. Zellers [12, 13] investigated pitch and segmental lengthening in relation to turn boundaries, and found that variation in phrase-final lengthening was a better predictor of turn hold than variations in pitch. In fact, pitch variations in the stimuli only influenced listeners' judgments if there was no length difference between the two stimuli they heard.

Zellers [12, 13] investigated two types of pitch variation: variation in the height of a final H peak in a turn, and variation in the level at which the final pitch fall ended (i.e. whether it fell all the way to the speaker's baseline or ended higher). A non-baseline end to the final pitch fall only contributed to listeners' judgments about turn change or turn hold in conjunction with variation in the height of the final pitch peak. However, such "truncated" pitch falls are not common in Swedish, and Bruce [5] suggests that they are to be associated phonologically with intermediate phrase boundaries rather than intonation phrase-final boundaries. However, Bruce [5] points out that, at least in read texts, a rising boundary tone (i.e. LH%) is possible in addition to the more common low boundary tone (i.e. L%). He associates this rising boundary tone with the English "continuation rise", and although his discussion of it is limited to read text, it is not difficult to make the connection between this and the turn-hold-related pitch rises reported for English and Dutch, among other languages. The current experiment therefore tests whether variations in the final pitch contour shape (falling versus falling-rising) as well as variation in the size of this falling-rising pitch "tail" contribute to Swedish listeners' judgments about whether a speaker wants to continue speaking. In light of previous results reported for Swedish that flat intonation, but not necessarily rising intonation, is associated with turn hold, the current study also investigates whether the height of the final tail rise influences listeners' judgments about turn hold.

Previous studies in German (Schneider & Lintfert [14]; Ambrazaitis [15]; Schneider et al. [16]) have suggested that listeners are very sensitive to differences in the size of final pitch tail rises, allowing for possible identification of three phonological categories discriminated in part by the final pitch height (cf. Ambrazaitis [15]). However, the research reported above suggests that only two phonological categories are necessary for Swedish phrase-final boundaries. The current study therefore also investigates Swedish listeners' ability to distinguish between different pitch contours independently of any functional question.

2. Experiment design

2.1. Stimuli

The stimuli for this experiment were created at the same time as those used by Zellers [12, 13]. They consisted of conversational turns taken from the DEAL corpus (Hjalmarsson et al. [17]), in which participants role-played bargaining at a flea market. The turns chosen for the current experiment had to be recognizably syntactically complete, with an accent on the final word but not the final syllable, and with a pitch contour consistent with the presence of a final focal accent (i.e. an F0 peak). In addition, for purposes of the pitch resynthesis, the segmental structure in the final word was restricted as much as possible to voiced segments. Three turns were used as the base tokens for the current experiment:

- *Designen är svensk men tillverkningen [är] i Kina.* (The design is Swedish but the manufacturing [is done] in China.)
- *Nej, jag har en lite större.* (No, I have a slightly bigger one.)
- *Ja jag har tittat lite på ahm på den där sågen.* (Yes I've looked a little at uhm at that saw.)

The *Kina* and *större* turns were produced by the same male speaker, while the *sågen* turn was produced by a female speaker. Both were native speakers of Central Swedish.

2.1.1. Resynthesis

Twenty-six versions of each sentence were created using PSOLA resynthesis in Praat [18]. Of these, 15 were used for the current study; these will be described here. The final word of each turn had a number of pitch contours superimposed. The pitch contours involved two kinds of variation. The Peak set had the final F0 peak of the turn set at 3 semitones (st), 5st, and 8st above the speaker's baseline. These stimuli all ended in an F0 fall to the speaker's baseline which coincided with the offset of phonation. The location of the pitch peak was not moved from where it occurred in the original turn; only the height of the peak was varied.

From each Peak height, a Tail set was also created. In order to create the Tail stimuli, an additional pitch point was set at a timepoint two-thirds of the way between the time of the pitch peak and the time of offset of phonation. This point was set to the speaker's baseline pitch. Then, the pitch at the offset of phonation was varied, being set at 0, 3, 5 or 8st above the speaker's baseline. The pitch variations are illustrated in Figure 1.

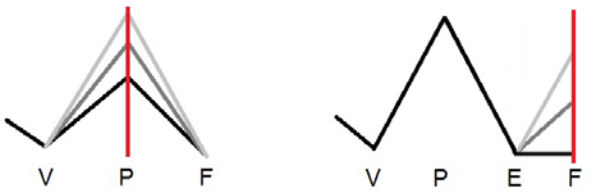

Figure 1: *Schematic pitch contours of resynthesized stimuli. Left: Peak height variations, with pre-final V(alley), pitch P(eak), and F(inal) low. Right: Tail variations, with V and P as in the pitch set, an E(lbow) at the speaker's baseline, and F(inal) pitch height at varying heights at or above the speaker's baseline.*

2.1.2. Naturalness ratings

In order to test whether the resynthesis was detectable and whether the resulting tokens were acceptable tokens in Swedish, 5 native speakers of Swedish listened to all of the resynthesized stimuli in randomized order and rated each token from 1 (very unnatural) to 4 (very natural). Since an exactly neutral rating would be 2.5, in the main experiment, only tokens which had a mean rating of 2.66 or higher were used. After excluding tokens which did not reach this threshold, 38 of the original 45 stimuli remained. (Note that a fourth base sentence was also included in the original resynthesis process, but so many tokens of this fourth sentence proved problematic in terms of naturalness that this sentence was excluded from the final experiment.)

2.2. Methodology

The experiment was presented using Praat MFC and consisted of two parts. In both parts, pairs of stimuli were presented that varied in peak height and end characteristics. The variation in end characteristics could be either a difference in the end contour shape (a full fall compared to a stimulus with a rising tail), or a difference in the height of the tail rise. Pairs were always two versions of the same sentence, so the only differences within each stimulus pair were prosodic ones.

In the first part of the experiment, hereafter the "Choose" task, listeners were asked to decide which version of the sentence sounded more like the speaker wanted to continue speaking after s/he had finished the turn. Listeners were also asked to indicate whether they were "ganska säker" (fairly sure) or "ganska osäker" (fairly unsure) about their response. In the second part of the experiment, hereafter the "Discriminate" task, listeners were asked only to indicate whether they heard a difference between the items in a pair.

Taking into account the fact that the acceptably natural stimuli were not balanced across the sentences, 159 total comparisons were made. Four versions of the experiment were created. The stimuli were divided into two chunks, A & B, consisting of 80 and 79 items respectively. In versions 1 and 2, listeners heard chunk A in the Choose task and chunk B in the Discriminate task; in versions 3 and 4 the reverse was the case. Versions 2 & 4 presented stimuli pairs in the opposite order to versions 1 & 3, in order to control for order-of-presentation effects. It was expected that listeners would be more sensitive in the Discriminate task, since simple discrimination would not necessarily have to rely on linguistic judgments, and pitch discrimination has been shown to be better in non-linguistic than linguistic stimuli, cf. e.g. Cummins et al. [19]. Thus all participants conducted the Choose task before the Discriminate task. Within each task, stimuli pairs were presented in a different random order for each participant, assigned by Praat. Listeners had the opportunity to listen to pairs a second time in the Choose task, but not the Discrimination task.

Twenty-four native speakers of Swedish (ages 18-39; 10 female), with no known hearing or language impairments, participated in the experiment. All were self-selected in response to a call for participants, and received a voucher for a cinema ticket in exchange for participation. The experiment took place in a quiet room in the Department of Speech, Music & Hearing, KTH. Most participants took about 30 minutes to complete the experiment.

2.3. Hypotheses

For the Choose task, in which listeners are asked which stimulus is more congruent with the speaker continuing to speak following the present turn,

- Higher pitch peaks will lead to a turn hold interpretation
- Tails ending above the speaker's baseline pitch may also contribute to a turn hold interpretation.

For the Discrimination task, in which listeners are asked whether pairs of stimuli are the same or different,

- Discrimination will be more sensitive to phonetic variation than the Choose task.

3. Results

3.1. Choose task

A logistic regression model, calculated using the lme4 package in R, shows a significant main effect of the height of the pitch peak, but no significant effects of either contour shape or tail end height, on listeners' choice of which stimulus in a pair sounded more like the speaker was going to continue speaking. The statistical model is shown in Table 1.

Although listeners did not rely on the tail contour characteristics for their judgments about finality, those characteristics did influence their certainty: specifically, there was a significant interaction between the height of the pitch peak and the height of the tail end. Statistics are shown in Table 2. The interaction seems to indicate that when both a large peak height difference (5st) and a large tail end height difference (5 or 8st) were present between the stimuli, listeners were more confident in their judgments.

Predictor	Estimate	SE(B)	e^B	z-score (p-value)
intercept	-0.128	0.207	0.880	-0.62 (0.54)
HeightDiff	-0.072	0.020	0.930	-3.66 (0.00***)
TailHtDiff	-0.007	0.012	0.993	-0.67 (0.54)
ContShape	0.088	0.207	1.092	0.43 (0.67)
Ht*TailHt	-0.002	0.004	0.998	-0.37 (0.71)
$\chi^2 = 13.995$, df = 4, $p < 0.01$**				

Table 1: *Logistic regression model for responses in the Choose task. Base sentence was included as a random factor in all statistical models.*

Predictor	Estimate	SE(B)	e^B	z-score (p-value)
intercept	0.309	0.242	1.362	1.27 (0.20)
HeightDiff	0.087	0.020	1.091	4.34 (0.00***)
TailHtDiff	0.032	0.211	1.033	2.70 (0.01**)
ContShape	-0.234	0.211	0.791	-1.11 (0.27)
Ht*TailHt	0.010	0.004	1.010	2.35 (0.02*)
$\chi^2 = 122.31$, df = 3, $p < 0.001$***				

Table 2: *Logistic regression model for goodness judgments in the Choose task.*

3.2. Discrimination task

For the Discrimination task, a logistic regression gives significant main effects for a difference in peak height, difference in contour shape, and difference in the height of the end of the tail, as well as a significant interaction between the difference in peak height and the difference in the height of the end of the tail, as predictors for whether participants could identify the two stimuli as different. The statistical model is shown in Table 3.

Predictor	Estimate	SE(B)	e^B	z-score (p-value)
intercept	1.256	0.225	3.510	5.59 (0.00***)
HeightDiff	-0.514	0.080	0.598	-6-41 (0.00***)
TailHtDiff	-0.133	0.046	0.875	-2.92 (0.00**)
ContShape	-0.652	0.308	0.521	-2.12 (0.03*)
Ht*TailHt	0.033	0.016	1.033	1.998 (0.04*)
$\chi^2 = 252.44$, df = 3, $p < 0.001$***				

Table 3: *Logistic regression model for judgments in the Discrimination task.*

A visual inspection of the data suggests that the interaction between the peak height and the tail end height as cues for discrimination is based around the peak height being a stronger cue. When the peak heights differed by 3 or 5st, a difference in tail height did not improve discriminability. However, when the peak heights were the same or differed by 2st, large differences in tail height (5 or 8st difference) improved discriminability of the stimuli. These discriminability differences are shown in Figure 2.

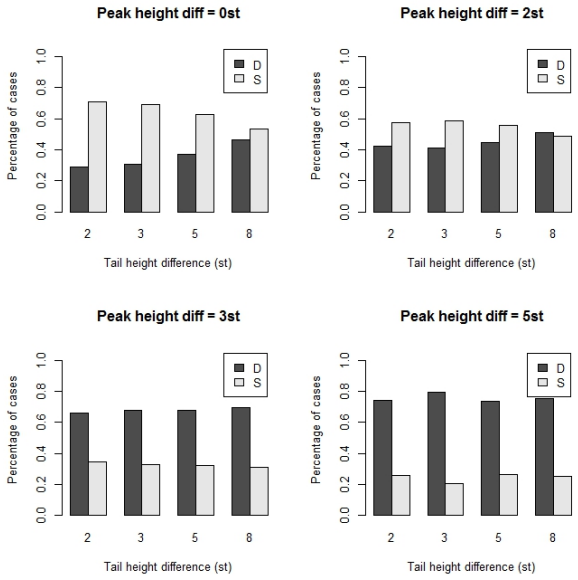

Figure 2: *Effect of tail height differences on Discrimination under different peak height difference conditions. Bar color indicates discrimination responses: D = different, S = same.*

4. Discussion

Of the three hypotheses presented in section 2.3, the first and third appear to gain support from the current data, while the second does not. As found by Zellers [12, 13], higher F0 peaks contribute to listener judgments that a speaker is planning to continue speaking. However, for Swedish listeners, variation in F0 tail height does not contribute to listener judgments about whether the speaker intends to continue speaking, despite the fact that (at least some of) these differences are

auditorily discriminable to listeners both in the Choose and in the Discriminate tasks.

4.1. Differences between Choose and Discriminate

A comparison between the results from the Discriminate task and the certainty judgments in the Choose task suggests, unsurprisingly, that listeners adopted different listening strategies in the two tasks. Due to the different nature of the responses, it is not possible to make an exact comparison across the two tasks; however, a few patterns in the responses stand out.

While "Uncertain" responses in the Choose task could indicate that there was no functional difference between the two stimuli listeners heard even when there was an audible prosodic difference, it is clear from the statistical model reported in Table 2 that prosodic factors also had an influence on how certain listeners were about their responses. Specifically, when pairs were more prosodically different, listeners were more confident in their judgments, even though they apparently did not take the height of the pitch tail end into account in terms of their judgment about turn hold signaling. However, in order for the simple presence of prosodic difference between the stimuli to influence certainty, this difference had to be drastic: a large height difference combined with large tail end differences improved certainty judgments, but large tail end differences apparently did not improve certainty with smaller height differences.

In contrast to the relatively limited influence of tail height on responses in the Choose task, tail height had a stronger effect in the Discriminate task. When the difference in pitch peak heights between pairs of stimuli in this task was nonexistent or difficult to discriminate (i.e. 0 or 2 st difference), listeners drew on differences in the tail end heights in order to be able to discriminate between stimuli. However, when the difference between pitch peak heights was larger, a difference in tail heights did not improve discriminability. Thus, although sensitivity to tails was apparently greater in the Discriminate task than in the Choose task (even accounting for the difference in reporting), this sensitivity appears to have been somewhat differently structured in the two different tasks. Specifically, in the Choose tasks, listeners apparently prioritized a combination of cues, while in the Discrimination task the cues could be, and were, used individually. This in turn suggests that listeners applied different listening strategies in the two tasks, and that in the Choose task, which required a more functional interpretation of the stimuli, the prosodic information from the peak and the tail may have been treated in a more integrated fashion than in Discrimination, where the identification of any single difference was enough to affect judgments.

4.2. Functions of pitch tail variation

Although different tail heights were discriminable for the Swedish listeners, Ambrazaitis [20] reports that when this particular pitch contour occurs naturally in Swedish speech, it tends to have a rise of about 5st, and normally ends at a lower height than the focal F0 peak. In German, in comparison, the tail rise height is similar, but it tends to end above the final F0 peak (in this case a pitch accent). As mentioned in section 1, perceptual evidence suggests that there are possibly three phrase-final categories in German: a fall, a fall with a slight rise, and a rise (Ambrazaitis [15]; cf. also Schneider & Lintfert [14], Schneider et al. [16]). These final categories have been argued to be related to turn-taking in that they indicate differing degrees of finality in what has been said.

Zellers [12, 13] suggests that the reason Swedish listeners prefer stimuli with higher F0 peaks for signaling turn hold is not directly related to a system of turn-taking cues, but rather to prominence. If direct turn-taking cues (e.g. lengthening) are not available, listeners take information from the sentence pragmatics: higher, more prominent peaks are associated with topicalization, and topicalizing a referent would seem to imply that there is more to be said about it. Gussenhoven [21] suggests that boundary tone height does not contribute to perceived prominence because it does not influence a hypothesized pitch reference line. The fact that the listeners essentially ignored the tail variations even though they were in many cases discriminable suggests that this may be a plausible interpretation of the listeners' behavior. This explanation leaves the improvement in certainty judgments related to larger prosodic differences unaccounted for, however. It is possible that certainty was improved not directly by the prosodic characteristics themselves, but rather by the particularly easily perceptible difference between the two stimuli, which simply allowed listeners to be confident that the two stimuli were different. Since all of the participants were also highly proficient second-language speakers of English, they may also have been susceptible to influence from prosodic cueing in that language.

The fact the tail variation did not have much influence on decisions in the Choose task does not, of course, imply that such pitch tails do not have any function in Swedish. Ambrazaitis [20] points out that final fall-rises are often used in "request addresses", but acknowledges that little else has been said about the function of this contour in Swedish. The current study suggests, however, that the classification of a final rising contour in Swedish as a "continuation rise" (cf. Bruce [5]) requires more refinement, since it apparently does not lead listeners to assume that a speaker's turn will continue. Bruce indicates that the rising contour is more common in read speech, so it may be that a final rise in Swedish is more appropriate in formal linguistic situations than casual ones.

5. Conclusions

The current study investigated the role of pitch contour size and shape on Swedish listeners' perception of turn hold. While peak height contributes to this perception, the end height of a pitch tail apparently does not, contrary to reports for a number of other languages, and despite the fact that variation in tail end height was perceptible in a discrimination task. This lends support to the idea that turn-transition cueing has language-specific aspects; further research is necessary to determine the extent to which a "continuation rise" may be relevant in other contexts in Swedish.

6. Acknowledgements

I am very grateful to David House for guidance on the experimental design, to Anna Hjalmarsson for providing me with her stimuli, to Niklas Vanhainen for proofreading my Swedish, and to Oliver Niebuhr for pointing me in a helpful direction for interpreting the data. This research was supported by the postdoctoral grant "Perception of prosody in linguistic contexts" (VR-435-2011-6871) from the Swedish Research Council (Vetenskapsrådet).

7. References

[1] Local, J.K., Kelly, J. & Wells, W.H.G. (1986) Towards a phonology for conversation: turn-taking in Tyneside English. *Journal of Linguistics* 22: 411-437.

[2] Gravano, A. & Hirschberg, J. (2009) Turn-yielding cues in task-oriented dialogue. *Proceedings of SIGDIAL 2009*, Queen Mary University of London, UK, 253-261.

[3] Caspers, J. (2003) Local speech melody as a limiting factor in the turn-taking system in Dutch. *Journal of Phonetics* 31: 251-276.

[4] Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A. & Den , Y. (1998) An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. *Language & Speech* 41: 295-321.

[5] Bruce, G. (1998) *Allmän och svensk prosodi. Praktisk linvistik 16.* Lunds Universitet.

[6] Gårding, E. (1989) Intonation in Swedish. *Lund University Department of Linguistics Working Papers* 35: 63-88.

[7] Gussenhoven, C. & Bruce, G. (1999). Word prosody and intonation. In van der Hulst, H. (Ed.) *Word prosodic systems in the languages of Europe.* Berlin/New York: de Gruyter, 233-271.

[8] House, D. (2004) Final rises and Swedish question intonation. In *Proceedings of FONETIK 2004*, Stockholm University, Sweden, pp. 56-59.

[9] House, D. (2006) Perception and production of phrase-final intonation in Swedish questions. In Bruce, G., & Horne, M. (Eds.), *Nordic Prosody, Proceedings of the IXth Conference, Lund 2004*. Frankfurt am Main: Peter Lang, 127-136.

[10] Hjalmarsson, A. (2011) The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication* 53: 23-25.

[11] Edlund, J. & Heldner, M. (2005) Exploring prosody in interaction control. *Phonetica* 62: 215-226.

[12] Zellers, M. (2013) Pitch and lengthening as cues to turn transition in Swedish. In *Proceedings of 14th Interspeech*, Lyon, France.

[13] Zellers, M. (submitted) Prosodic variation for turn transition in Swedish.

[14] Schneider, K. & Lintfert, B. (2003) Categorical Perception of Boundary Tones in German. In *Proceedings of 15th ICPhS*, Barcelona, Spain.

[15] Ambrazaitis, G. (2005) Between fall and fall-rise: substance-function relations in German phrase-final intonation contours. *Phonetica* 62: 196-214.

[16] Schneider, K., Dogil, G. & Möbius, B. (2009) German boundary tones show categorical perception and a perceptual magnet effect when presented in different contexts. In *Proceedings of Interspeech 2009*, Brighton, England. 2519-2522.

[17] Hjalmarsson, A., Wik, P. & Brusk, J. (2007) Dealing with DEAL: a dialogue system for conversation training. In *Proceedings of SIGDIAL*, Antwerp, Belgium, 132–135.

[18] Boersma, P. & D. Weenink (2013). Praat: doing phonetics by computer [Computer program]. Available http://www.praat.org/

[19] Cummins, F., Doherty, C., & Dilley, L. (2006) Phrase-final pitch discrimination in English. In *Proceedings of Speech Prosody 2006,* Dresden, Germany. 467-470.

[20] Ambrazaitis, G. (2008) On final rises and fall-rises in German and Swedish. In Eriksson, A. & Lindh, J. (Ed.) *Proceedings FONETIK 2008*, Department of Linguistics, University of Gothenburg. 81-84.

[21] Gussenhoven, C. (2002) Intonation and interpretation: Phonetics and Phonology. In *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France. 47-57.